This document is downloaded from the
VTT's Research Information Portal
https://cris.vtt.fi

VTT Technical Research Centre of Finland

Dataset on researcher's mobility by utilising bibliometric data: Case of Finland-based scholars for the duration of 2008-2018

Hajikhani, Arash; Suominen, Arho

Published in:
Data in Brief

DOI:
10.1016/j.dib.2021.106764

Published: 01/02/2021

Document Version
Publisher's final version

License
CC BY

Link to publication

Please cite the original version:
Hajikhani, A., & Suominen, A. (2021). Dataset on researcher's mobility by utilising bibliometric data: Case of Finland-based scholars for the duration of 2008-2018. *Data in Brief*, *34*, 106764. [106764]. https://doi.org/10.1016/j.dib.2021.106764

VTT
http://www.vtt.fi
P.O. box 1000FI-02044 VTT
Finland

By using VTT's Research Information Portal you are bound by the following Terms & Conditions.

I have read and I understand the following statement:

This document is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of this document is not permitted, except duplication for research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered for sale.

Download date: 05. May. 2023

Data in Brief 34 (2021) 106764

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset on researcher's mobility by utilising bibliometric data: Case of Finland-based scholars for the duration of 2008-2018

Arash Hajikhani[a,*], Arho Suominen[a,b]

[a] Quantitative Science and Technology Studies, VTT Technical Research Centre of Finland, Tekniikantie 21, 02044 Espoo, Finland
[b] Industrial Engineering, Tampere University, Korkeakoulunkatu 8. PL 541, 33014 Tampereen yliopisto, Finland

ARTICLE INFO

Article history:
Received 3 November 2020
Revised 8 January 2021
Accepted 13 January 2021
Available online 16 January 2021

Keywords:
Bibliometrics
Scientometrics
Researchers mobility
Data science

ABSTRACT

Scientific discoveries are the result of global collaboration and often the multidisciplinary nature of collaborations. A core element of these successful collaborations will materialise through a researcher's mobility in location and disciplinary focus. Researchers experience numerous opportunities to practice locational mobility throughout their careers as well as by conducting multidisciplinary research. Both changes have short- and long-term impacts on individual researchers and science, technology, and innovation systems that have an immediate interest for the public and private research and development funding mechanisms. With the advancement in data science tools and increasing computational capacities, we can use bibliometric data for calculating a researcher's mobility on location and a disciplinary focus over time. We looked at Finland as a case, and by incorporating analytical procedures, the processed data is capable of delivering insights on researcher mobility between cities over time as well as disciplinary change over time. This dataset can reveal hidden dynamics in the scholar's career progress. If combined with funding information and mission-oriented policies, the dataset can evaluate the long-lasting effect of instruments in mobilising researchers, steering research agendas, and consequently the resulting impacts.

* Corresponding author.
E-mail address: arash.hajikhani@vtt.fi (A. Hajikhani).
Social media: (A. Hajikhani)

https://doi.org/10.1016/j.dib.2021.106764
2352-3409/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

© 2021 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/)

Specifications Table

Subject	Social Sciences, Library and Information Science
Specific subject area	Bibliometrics, Scientometrics, Researchers mobility, Disciplinary mobility
Type of data	Datasets, XLSX files, Graph network data
How data were acquired	Clarivate Web of Science Core Collection. Calculated over the publication's bibliometric data to model the change of location within the researcher's course of scholarship as well as the researcher's disciplinary change. Jupyter Notebooks and Python scripts were used to analyse, filter, clean and process the data.
Data format	Analysed, Triangulated, Cleaned, Filtered
Parameters for data collection	All scientific publication data with at least one of the authors were based in Finland from 2008 to 2018.
Description of data collection	The raw bibliometric data has been downloaded from the Web of Science Core Collection. The Python programming language was used to parse the bibliometric data for extracting the author's publication, their affiliation and their disciplinary focus at the time.
Data source location	Primary data source: Clarivate's Web of Science Core Collection
Data accessibility	Available at: http://dx.doi.org/10.17632/3hhdwz56c8.1

Value of the Data

- The data set consists of over 9 million location mobility events from 300 thousand individual researchers based in Finland for over 10 years. Thus, it can be used to analyse different patterns, such as the research funding impact, thematic funds impact, change of research focus and the long- and short-term effects of funding and mobility in researchers' careers.
- New variables are created. 'Location Mobility' counts the event if the scholar affiliation has been changed when time progresses in the research career. Also, the 'From' and 'To' column that provides the outgoing and incoming location on the city level. 'Disciplinary Change' captures the change of disciplinary focus of each scholar over time and generates the 'From' and 'To' columns that provide the changing disciplines.
- Researchers in different fields of knowledge can use these datasets to analyse the dynamics of Science Technology and Innovation (STI) considering the researcher's location mobility and disciplinary change.
- The dataset can inform private and public funding agencies on the impact of their funding and research steering on a researchers' career path and the general STI system. This way, it would be possible to design effective strategies for steering research & development funding activities and spot the various funding effects on multidisciplinary extensions of research outcomes.
- This dataset if combined with funding data, will allow researchers and policymakers to obfuscate the motivations of individual researcher mobility, ultimately clarifying the impacts of mobility to individual researchers and the research system.
- The computational procedure for compiling the data presented in this paper can be used as a guide for conducting similar studies around the globe.

1. Data Description

The dataset made available in this paper consists of identified Finland-based scholar's mobility (location and disciplinary focus) over 10 years (2008–2018). We still know relatively little

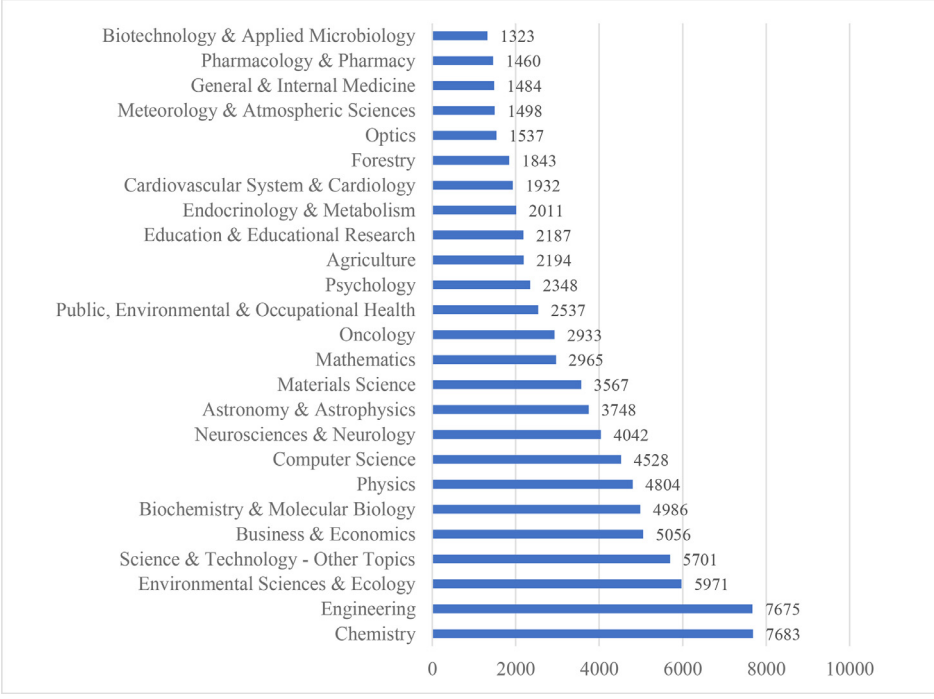

Fig. 1. Top 25 Finland's scientific activity distribution over WoS categories.

about the motivations of individual researcher mobility [1]. The literature suggests that mobility is more of a necessity than a choice to advance a research agenda [2]. We also know that a relatively large share of researchers has had a negative experience during their mobility period [3]. However, science policy often emphasises the role of mobility, and it is often embedded in funding and tenure decisions as an essential element. The policy looks towards mobility to improve the quality of science in science systems [4] while also integrating science systems such as the European Research Area [5]. This dataset will allow researchers and policymakers to obfuscate individual researcher mobility motivations within a location and disciplinary focus, ultimately clarifying the impact of mobility on individual researchers and the research system.

Clarivate Analytics Web of Science (WoS) is selected as the data source for compiling the dataset. We consider the WoS to be more appropriate for our compiled data set for several reasons. First, as a global citation database and comprehensive platform with over 159 million records and over 1.7 billion cited references, the WoS can track ideas across time and disciplines. Second, comparative and longitudinal studies have shown a consistent and reasonably stable quarterly growth for both publications and citations in the WoS Core Collection database [6]. Third, the WoS data structure is favourable because it consistently constructs a scholar's change in affiliation and subject category focus much more accurately than other indexing services. We also cross-check our research results with Scopus data, which extensively covers scientific publications. However, we refrain from using the new Google Scholar archiving services, as the accuracy of its citation counts has been seriously doubted [7–9].

We used 'Advance Search' in the Clarivate's Web of Science and selected the Core Collection database to search for all publication types between 2008 and 2018 with at least one author based in Finland. The descriptive analysis of Finnish science breadth and depth has been illustrated in Fig. 1 regarding publications volume distribution over scientific subject categories and the growing areas in Fig. 2. Finnish authors' scientific publication activities have then been

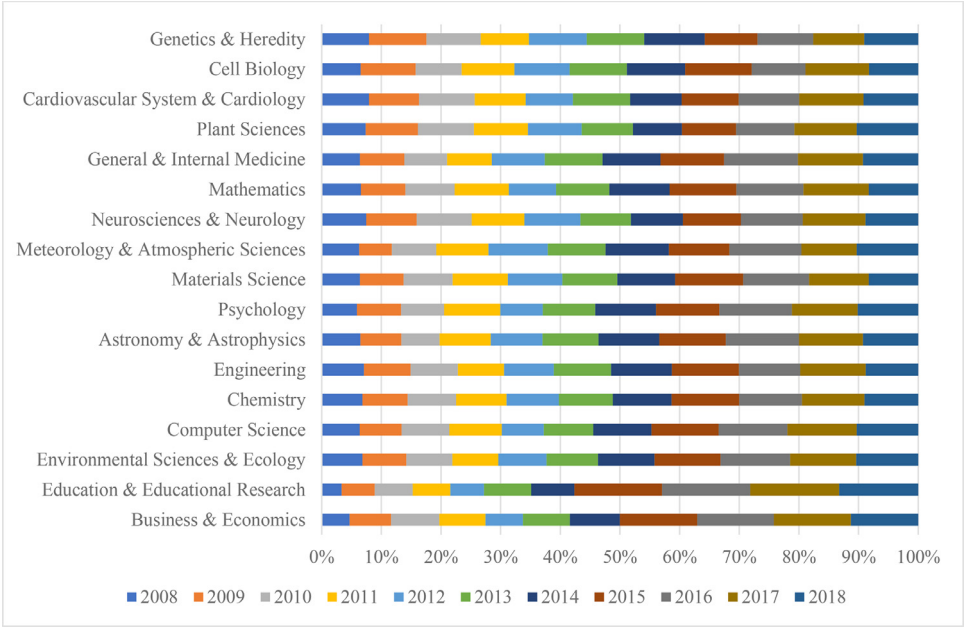

Fig. 2. Finland's top growing disciplinary categories for duration of 2008-2018.

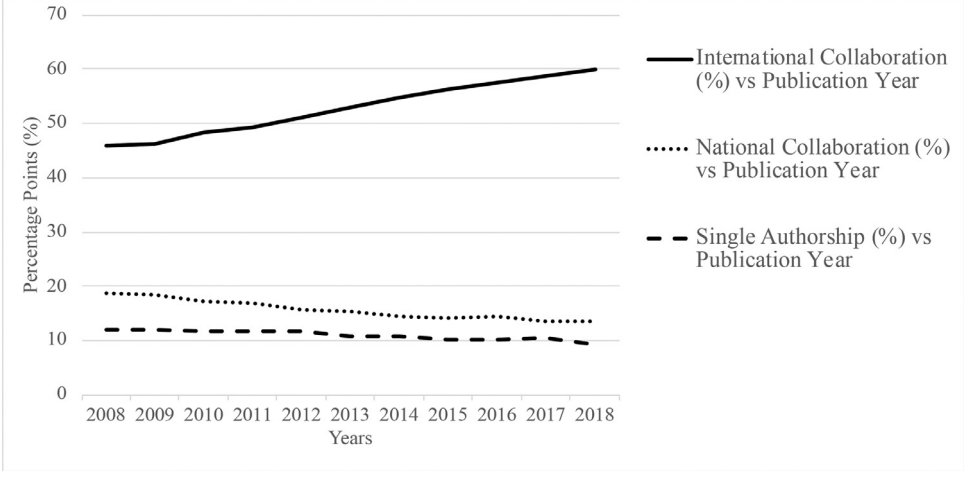

Fig. 3. Publication's type authorship benchmarking in Finland (2008-2018).

observed based on collaborations with international scholars, National collaboration and solo authorship for the study period, illustrated in Fig. 3. Furthermore, Finland's research excellence and collaboration activity have compared with EU 27 member states, illustrated in Fig. 4.

From the raw bibliometric data retrieved (data scheme illustrated in Fig. 5), a particular procedure took place to analyse, filter, clean and process the change in the authors' affiliation over their career trajectory while also picking up occasions where the disciplinary focus has changed in the authors' publication profile.

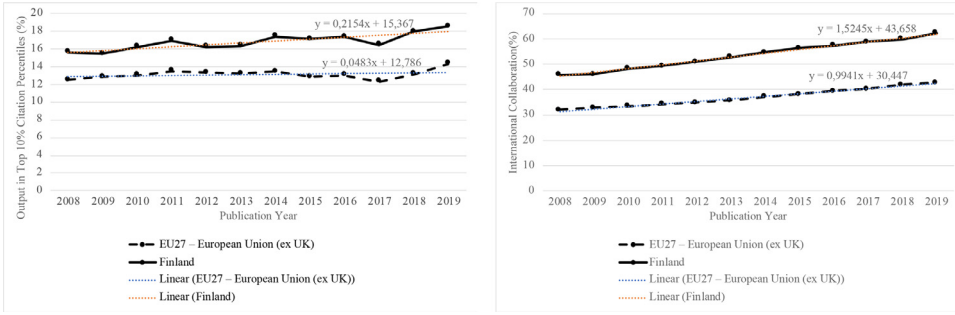

Fig. 4. Finland's research excellence and collaboration activity compared to EU27.

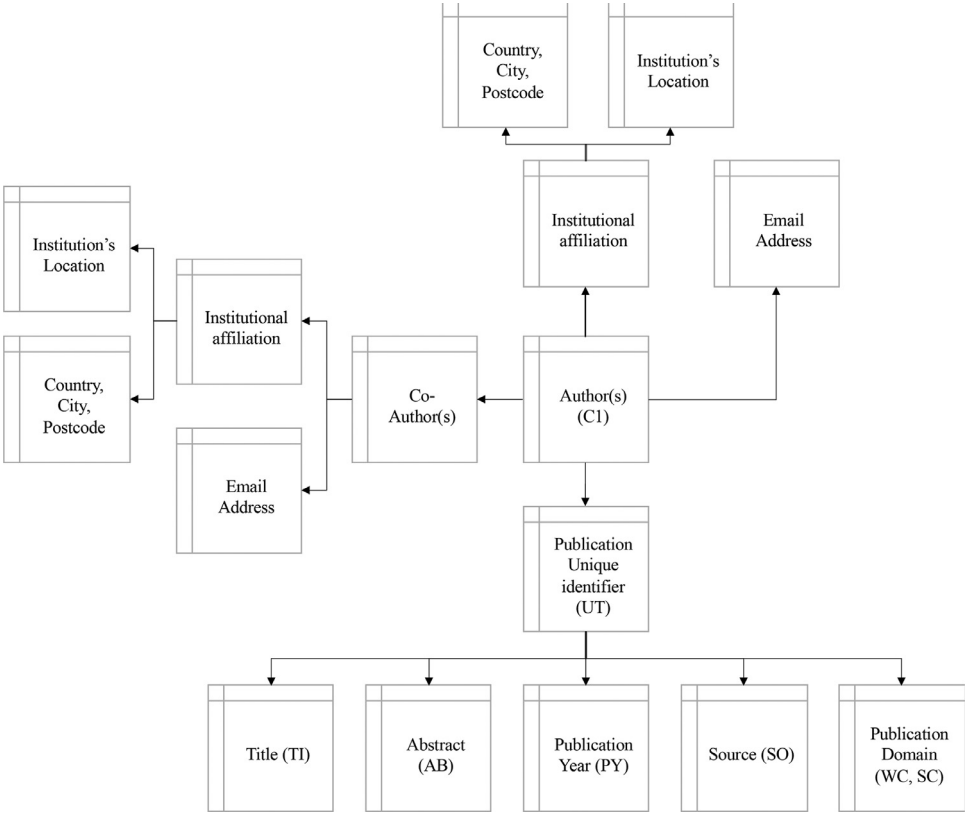

Fig. 5. WoS bibliometric data Authors field.

The final data set is provided in four individual XLSX files accessible from Mendeley Data at "http://dx.doi.org/10.17632/3hhdwz56c8.1". Table 1 presents the features contained in each of the specific data sets.

All four filtered and clean datasets are provided with examples for their visualisation and sensemaking. Section 2 while describes the methods and steps for producing the dataset, will show the network structure of location mobility of scholars in Fig. 6. Fig. 7 on the other hand, projects a snapshot of disciplinary change as a matrix illustration with numeric values for each

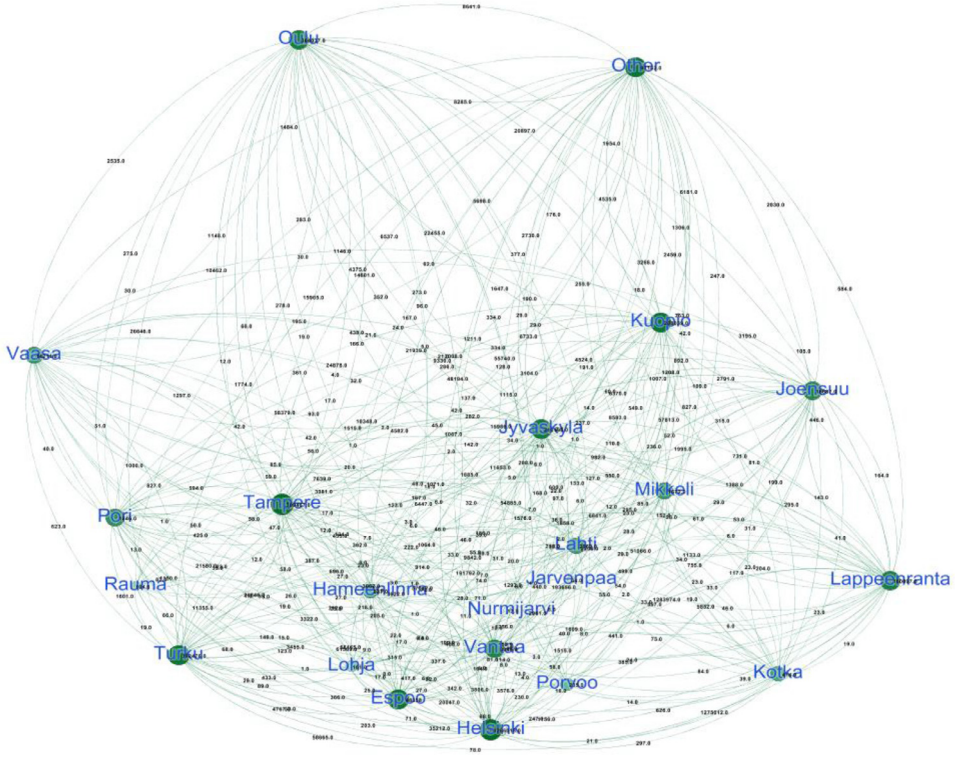

Fig. 6. Network visualisation of location mobility for period 2014–2018, Finnish scholars for the top 22 cities in Finland.

	Hematology	Oncology	Science & Technology - Other Topics	Neurosciences & Neurology	Genetics & Heredity	Cardiovascular System & Cardiology	Astronomy & Astrophysics	Meteorology & Atmospheric Sciences	Biochemistry & Molecular Biology
Hematology	8339	313	89	16	46	118	4	0	55
Oncology	291	133389	2008	117	2008	129	20	8	955
Science & Technology - Other Topics	76	1982	355292	1180	3387	797	695	467	1693
Neurosciences & Neurology	22	120	1082	102123	645	250	9	0	327
Genetics & Heredity	37	1843	3471	665	324401	591	2	8	887
Cardiovascular System & Cardiology	102	125	762	306	686	66864	4	4	177
Astronomy & Astrophysics	2	20	885	10	8	0	1153346	69	25
Meteorology & Atmospheric Sciences	0	8	394	3	7	2	73	20701	11
Biochemistry & Molecular Biology	48	1060	1720	385	991	199	23	7	101997
Audiology & Speech-Language Pathology	0	4	17	46	1	2	0	0	2
Entomology	2	4	155	1	0	0	1	0	18
Chemistry	7	36	1069	30	12	17	28	88	589
Cell Biology	17	181	427	125	130	73	8	1	312
Geology	1	1	127	5	2	2	132	288	2
Computer Science	0	15	177	59	18	10	9	15	74
Surgery	14	56	23	32	8	88	18	2	8
Engineering	7	46	572	88	12	20	68	200	121
Psychiatry	2	17	148	270	70	27	8	1	43
Business & Economics	1	5	165	21	5	8	2	2	7
Biotechnology & Applied Microbiology	13	121	363	30	73	60	6	5	258
Microbiology	1	11	187	9	28	2	1	3	109
Food Science & Technology	7	71	267	8	38	9	4	2	63
Fisheries	0	0	19	0	8	1	2	1	8
Biophysics	22	32	67	25	3	7	6	8	57
Forestry	0	2	122	1	3	1	2	14	17
Immunology	44	117	204	88	38	26	3	2	132
General & Internal Medicine	51	856	536	485	338	721	27	4	170

Fig. 7. Disciplinary change matrix illustration.

Table 1
Dataset specifications.

Dataset name	Rows/Columns	Sheets name
Author_Mobility_Matrix	Rows: City names (From) Columns: City names (To)	Sheet1: Location Mobility for 2008–2018
Author_Mobility_Network	Column1: City name (From) Column2: City name (To) Column3: Accumulative times of mobility	Sheet2: Location Mobility for 2009–2013 Sheet3: Location Mobility for 2014–2018
Discipline_Mobility_Matrix	Rows: Disciplinary category (From) Columns: Disciplinary category (To)	Sheet1: Disciplinary Mobility for 2008–2018
Discipline_Mobility_Network	Column1: Disciplinary category (From) Column2: Disciplinary category (To) Column3: Accumulative times of mobility	Sheet2: Disciplinary Mobility for 2009–2013 Sheet3: Disciplinary Mobility for 2014–2018

discipline category change. Fig. 8 illustrates the disciplinary change within a network structure where nodes are fields and edges are the mobility among them.

2. Experimental Design, Materials and Methods

2.1. Data collection

Based on the search query initiated in the WoS core collection, we captured over 130,000 bibliometric full records that cover the Journal category of publications between 2008 and 2018 (November 27) in which at least one of the authors has been based in Finland (CU = Finland). We retrieved the data in batches of 500 records as CSV (comma-separated values) files and combined them into one file for easy loading and processing of the raw data. The retrieved raw bibliometric data has then been processed by Jupyter Netbook and Python[1] version 3.7.6 for further parsing, filtration and triangulation of the data. A quick observation of the data describing state of the art in Finnish science indicates Finnish science's diversity, which covers approximately 160 Web of Science subject categories. From a disciplinary perspective, Finland-based scholars' top 25 most contributed categories and the publications output counts associated with the categories are illustrated in Fig. 1.

Detailed analysis of each subject category publications counts for each year; we could identify the top growing and accelerating subject categories over the past 10 years. Fig. 2 shows the top fields representing a positive average growth in term of publication outputs for the duration of 2008 to 2018.

Over the 10 years of scientific activity, Finland-based authors have conducted over 9.5 million acts of authorship (by 340 thousand unique authors) with over 270 million acts of authorship offered by international collaborators. The numbers are extracted by isolating each author per publication and controlling for authors' uniqueness in case of unique authorship contribution. Authorship can be translated as one's effort in establishing a scientific publication, which in Finland's case, is carried out by 339,918 unique authors from 2008 to 2018. Fig. 3 shows a trend line of international collaboration, national collaboration and solo authorship within Finland over the years.

The trend lines in Fig. 3 confirm the fact that Finnish science is becoming more collaborative, as single authorship is reducing also the international collaboration has increased by 15% while in contrast, national collaboration has decreased by 7% over the 10 years of the dataset's coverage. Finnish research activity on quality of publication outputs and international collaborations

[1] Python is an interpreted, high-level and general-purpose programming language. https://www.python.org

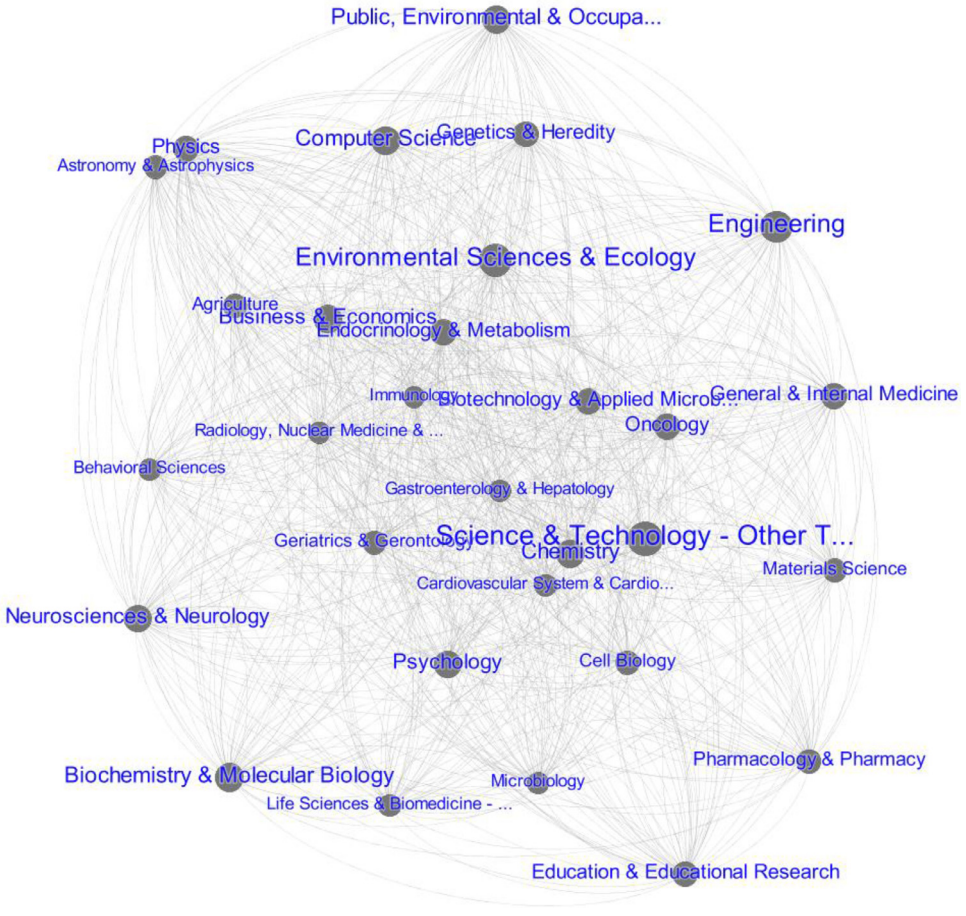

Fig. 8. Network visualisation of disciplinary change.

is benchmarked with other EU member states for comparative reasons. Fig. 4 illustrates the publication outputs of top 10% citation percentile over the years (Left side). On the right side, the international collaboration percentage of Finland is put into perspective with the average 27 EU members.

It is evident from Fig. 4 that Finland performs higher than the EU27 average in some of the research excellence metrics. For output in top journals with a measure of two percentage point deference with EU27 and higher growth rate by looking at the linear projection of publication's activity trend. International collaboration comparison shows Finland outperforming average EU27 by almost 20 percentage points with a slightly higher growth trajectory in perspective.

2.2. Author's affiliation extraction

Our search query identified Finland as the country in the 'CU' field tag and specified the 10 years to be from 2008 to 2018. The query was constructed, and results were retrieved on 20.11.2018. Filtering for the Article type of documents, 131,632 records were identified.

Table 2
Location mobility construction.

Name	Year	Address	Changed Year	Changed Address
Oliver	2018	Lappeenranta	0	0
Oliver	2018	Lappeenranta	0	0
Oliver	2019	Lappeenranta	1	0
Oliver	2019	Oulu	0	1
Oliver	2019	Tampere	0	1
Emily	2018	Germany	0	0
Emily	2019	Germany	1	0
Emily	2019	Germany	0	0
Emily	2020	Helsinki	1	1
Emily	2020	Germany	0	1

The retrieved raw bibliometric data will be first utilised to identify and analyse Finland-based researchers' affiliation and change of affiliation over time as a proxy for mobility. The analysis process reads the tabulator csv-delimited files and extracts the author's Publication unique identifier (UT), Publication year (PY), Publication Domain (WC, SC) and Authors (C1) fields for supplementary analysis. The authors' names and affiliations listed for each publication are separated into the single author's name column and affiliations to the organisation column. Fig. 5 shows the relational structure of metadata available for the Authors field.

For each author of a paper, the data structure stores a list of co-authors. Each author is also linked to an affiliation, if available. The script separates the authors' names and their affiliation into different columns. If the number of authors/affiliations matches the number of authors in the AF field, then the abbreviated authors' names are expanded to the full names in the Authors field. The scripts break down the affiliation string into pieces as each address information type is separated with a comma (','). Finland-based authors have collaborated extensively with international co-authors. As we are interested in Finland-based authors' mobility, we filter the data to rows where the string has 'Finland' in the affiliation address. Finally, each author's records are linked to the records' publication id, publication years, title, and subject category.

2.3. Authors mobility

The script is operationalised to group all the rows with the same author name and then sort the publication activity by the publication year in ascending order. This triangulation of the data lets us know the author's publications over the years—considering their affiliation that contains a physical address. The script then picks up if a change in an address string value occurs comparing it to the previous row and registers it in a separate row, 'Changed_Address'. If the row's publication year number changes, it identifies it in the 'Changed_Year' column. If there is no change, then it is identified as zero. As a simplistic example, Table 2 illustrates the created metadata once the authors, publication year and affiliation are generated. The table shows Authors' Name', Publication' Year' and affiliation 'Address' where the computational procedure has added the columns' Changed_Address' and 'Changed_Year' to identify the change of location over the years of scholarly publication activity.

The added value to the data structure will enable various ways of reporting the data. For example, referring to Table 2, we can infer that Oliver changed years '1' time and changed locations '2' times. Emily changed years '2' times and changed locations '2' times. The total changed addresses for the year 2019 is '2' times. The mathematical annotation of the author address mobility calculations can be seen in formulas (1). *X* refers to authors location for a range of all

Table 3
Location mobility matrix illustration.

From / To	Germany	Tampere	Helsinki	Lappeenranta	Oulu
Germany	2	2	2	2	0
Tampere	1	0	1	0	0
Helsinki	2	1	1	0	0
Lappeenranta	0	0	0	2	1
Oulu	0	1	0	0	0

authors (a_m) in all years (i_n) and the location mobility is annotated as $\Delta X_{i_n}^{a_m}$.

$$X_{i_n}^{a_m} = \begin{bmatrix} X_{i_1}^{a_1} & \cdots & X_{i_n}^{a_1} \\ \vdots & \ddots & \vdots \\ X_{i_1}^{a_m} & \cdots & X_{i_n}^{a_m} \end{bmatrix}$$

$$i \rightarrow (i_1, i_2, i_3, \ldots, i_n)$$

$$a \rightarrow (a_1, a_2, a_3, \ldots, a_n) \tag{1}$$

$$\Delta X_{i_n}^{a_m} = \begin{cases} 0, & if\ X_{i_n}^{a_m} = X_{i_{n+1}}^{a_m} \\ 1, & if\ X_{i_n}^{a_m} \neq X_{i_{n+1}}^{a_m} \end{cases}$$

$$All\ scholar's\ overall\ mobility\ in\ years = \sum X_{i_n}^{a_m}$$

2.4. Location mobility network graph

The most property of the created metadata on location change and the year change for each author is trough aggregation (either location or year). This will enable us to observe the data for top outgoing and incoming locations over the years or for specific periods. In other words, to get the network graph matrix (adjacency matrix) is to see the total changes between addresses—for example, how many times scholars moved from 'Germany' to 'Helsinki' in 2018. Table 3 is an example of an illustration in which the total change between locations is visible in one matrix.

For visualisation purposes, we compute the network structure of the adjacency matrix. The results can be exported in a format that can be read by network visualising software, such as Gephi[2]. Fig. 6 illustrates the authors' mobility data covering 10 years among the top 22 cities in Finland.

Each node represents a city, and outgoing or incoming links to cities indicate scholars' mobility over time, which is written on the edge as numerical values. Each city has an edge to itself that indicates the number of times the mobility did not take place, and authors stayed in their original location.

2.5. Discipline mobility

The same analysis for calculating author mobility can be constructed for the author's change of subject category in the order: From subject category To subject category over the years. For

[2] Gephi is open-source network analysis and visualization software. https://gephi.org/

each record of bibliometric publication, there is metadata on the subject category of the publication. WoS core collections offer two classification regimes, the Web of Science Category (WC) and Subject Categories (SC), which offers over 252 different existing categories [10]. Each publication has either one subject category or multiple categories, separated by a semicolon (;). In this exercise, we concentrated on SC first-level categories. The same analytical process has been replicated in the discipline of mobility data curation with the difference that instead of the city location name, SC are the objects where a change in them over time is captured in the analytical process. This process's outcome can show the change in the author's disciplinary focus over the overall study time and in different periods. Fig. 7 illustrates an example; there total change between disciplinary categories is visible in one matrix.

In Fig. 7, the matrix of diagonal values indicates the changes between the same subject category; this is the largest number as scholars often tend to stay focused on their discipline. However, we tend to keep this value because it will be informative in relative measures if internal disciplinary consistency will be compared to another disciplinary category. For network visualisation purposes with software such as Gephi, another triangulation of the data is provided. Importing the 'Diciplinary_Mobility_Network.xlsx' sheet1, which covers 2014–2018, to Gephi could help us project the subject categories with the highest mobility rate among themselves. Fig. 8 illustrates the top 20% of SCs, where nodes represent an SC and edges indicate mobility.

Ethics Statement

The required permissions have been obtained from Clarivate for publishing the compiled dataset.

CRediT Author Statement

Arash Hajikhani: Conceptualisation, Methodology, Software, Validation, Investigation, Writing - Original Draft, Visualisation, Project administration; **Arho Suominen:** Conceptualisation, Validation, Writing- Reviewing and Editing

Declaration of Competing Interest

None.

Acknowledgement

This work is supported by Business Finland under the project called "Advanced methods for the impact assessment of technological and business experimentation (INNOPACT)".

References

[1] A. Fernández-Zubieta, A. Geuna, C. Lawson, What do we know of the mobility of research scientists and impact on scientific production, in: Glob. Mobil. Res. Sci., Elsevier, 2015, pp. 1–33, doi:10.1016/b978-0-12-801396-0.00001-6.
[2] S. Morano-Foadi, Scientific mobility, career progression, and excellence in the european research area1, Int. Migr. 43 (2005) 133–162, doi:10.1111/j.1468-2435.2005.00344.x.
[3] G. Melin, The dark side of mobility: negative experiences of doing a postdoc period abroad, Res. Eval. 14 (2005) 229–237, doi:10.3152/147154405781776102.
[4] L. Ackers, Moving people and knowledge: scientific mobility in the European Union, Int. Migr. 43 (2005) 99–131, doi:10.1111/j.1468-2435.2005.00343.x.
[5] H. Toivanen, A. Suominen, Epistemic integration of the european research area: the shifting geography of the knowledge base of finnish research, 1995-2010, Sci. Public Policy. 42 (2015) 549–566, doi:10.1093/scipol/scu066.

[6] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, E. Delgado López-Cózar, Google Scholar, Web of science, and scopus: a systematic comparison of citations in 252 subject categories, J. Informetr. 12 (2018) 1160–1177, doi:10.1016/j.joi.2018.09.002.
[7] A.W. Harzing, S. Alakangas, Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, Scientometrics 106 (2016) 787–804, doi:10.1007/s11192-015-1798-9.
[8] P. Jacsó, Metadata mega mess in google scholar, Online Inf. Rev. 34 (2010) 175–191, doi:10.1108/14684521011024191.
[9] M. Levine-Clark, E.L. Gil, A comparative citation analysis of web of science, scopus, and google scholar, J. Bus. Financ. Librariansh. 14 (2008) 32–46, doi:10.1080/08963560802176348.
[10] M. Boletta, New Web of Science categories reflect ever-evolving research - Web of Science Group, Clarivate Web Sci (2019) https://clarivate.com/webofsciencegroup/article/new-web-of-science-categories-reflect-ever-evolving-research/. Accessed October 25, 2020.